ProtGPT2 is a deep unsupervised language model for protein design

Year: 2022 | Citations: 669 | Authors: Noelia Ferruz, Steffen Schmidt, B. Höcker

---

Abstract

Protein design aims to build novel proteins customized for specific purposes, thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in Transformer-based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Here, motivated by this success, we describe ProtGPT2, a language model trained on the protein space that generates de novo protein sequences following the principles of natural ones. The generated proteins display natural amino acid propensities, while disorder predictions indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are distantly related to natural ones, and similarity networks further demonstrate that ProtGPT2 is sampling unexplored regions of protein space. AlphaFold prediction of ProtGPT2-sequences yields well-folded non-idealized structures with embodiments and large loops and reveals topologies not captured in current structure databases. ProtGPT2 generates sequences in a matter of seconds and is freely available. Protein design aims to build novel proteins customized for specific purposes, thereby holding the potential to tackle many environmental and biomedical problems. Here the authors apply some of the latest advances in natural language processing, generative Transformers, to train ProtGPT2, a language model that explores unseen regions of the protein space while designing proteins with nature-like properties.